

Vistrit Pandey

Chicago, IL • +1 215 452 9375 • vistritpandey@gmail.com • vistrit.space • linkedin.com/in/vistrit

AI Engineer • Applied AI • ML Systems and Inference • Forward Deployed Engineering

Experience

Blinkfire Analytics

Jan 2026 to Present

Software Engineer II, AI R&D

Chicago, IL

- Build production AI systems end to end across **model evaluation, inference, backend services, APIs, and customer facing applications**, working in Python, Go, TypeScript, Next.js, and Google Cloud.
- Design and operate local LLM inference infrastructure using **CUDA, TensorRT, NVIDIA Triton, Ollama, Docker, Kubernetes, and Linux**, optimizing model serving, latency, throughput, deployment control, and cost.
- Develop **reinforcement learning, synthetic data generation, fine tuning, and evaluation pipelines** over 12+ years of proprietary multimodal data for LLM training and benchmarking.
- Ship agentic and multimodal systems combining **RAG, semantic search, tool execution, Whisper, LLMs, and vision language models** for analytics, automation, audio understanding, and structured outputs.
- Partner with product, infrastructure, leadership, and enterprise stakeholders to take AI capabilities from research prototypes into production services and customer workflows.

Blinkfire Analytics

Apr 2024 to Dec 2025

Software Engineer I, AI R&D

Chicago, IL

- Shipped full stack AI and analytics products using **Python, Go, TypeScript, REST, gRPC, Redis, Next.js, Vue.js, Firebase, and GCP** across distributed service architectures.
- Integrated commercial and open weight models into semantic search, content intelligence, automation, and real time inference; built **fine tuning, quantization, serving, and inference optimization** workflows with PyTorch, TensorFlow, CUDA, TensorRT, and Triton.
- Built backend APIs, asynchronous processing systems, annotation tooling, and customer facing interfaces supporting production AI workflows.

Meettext (Independent Project)

Mar 2023 to Mar 2024

Founder and Software Engineer

Remote

- Founded and built an **AI communication platform** with Next.js, Python, Firebase, speech recognition, LLM APIs, and Stripe, scaling it to more than **20,000 users**.
- Led seven engineers and designers while owning system architecture, cloud deployment, payments, product delivery, user support, and technical execution from prototype through public launch.

Drexel University

Sept 2022 to Sept 2023

Course Assistant, Computer Science

Philadelphia, PA

- Mentored **500+ students** across algorithms, machine learning, computer vision, computational imaging, programming, and software engineering through labs, debugging, code reviews, and project evaluations.

Projects and Research

MachBoost | Open Source Local AI Infrastructure

machboost.xyz | [GitHub](https://github.com)

- Built a native macOS app, Python package, and team inference gateway for **resident MLX text and vision models**, with authenticated LAN serving, load aware routing, bounded queues, and OpenAI, Anthropic, and Ollama compatible APIs.
- Implemented repository aware prefix reuse, multimodal caching, and inference acceleration paths; measured **2.89x median speedup** across token identical repeated repository workloads under paired benchmarks.

Meettext | Forward Deployed Engineering Platform

meettext.xyz

- Built a deployment control plane for shipping **AI products, MCP servers, and APIs** into customer environments, with OpenAPI to MCP conversion, per customer capability isolation, versioned deployments, and multi destination delivery.
- Implemented OAuth and permission validation, security review gates, continuous health checks, incident diagnostics, credential lifecycle management, and end to end deployment verification.

MeetHealth | Applied AI and Model Evaluation

meethealth.us

- Built an AI platform for dietitians supporting 200+ organizations and generated **1M+ synthetic clinical diet plans** for guideline anchored evaluation, training data, and domain specific model development.

Education

University of Illinois Urbana Champaign

2026 to 2027 (Expected)

Master of Computer Science, Online

Illinois, US

Drexel University

2019 to 2023

Bachelor of Science in Computer Science, Minor in Information Systems

Pennsylvania, US

Technical Skills

Languages: Python, Go, TypeScript, JavaScript, SQL, Java, C++, Bash

AI and ML: PyTorch, TensorFlow, Hugging Face, MLX, RL, fine tuning, evaluation, synthetic data, quantization, RAG, multimodal AI, Whisper, VLMs

Systems and Product: CUDA, TensorRT, Triton, Ollama, Docker, Kubernetes, Linux, GCP, AWS, Redis, gRPC, React, Next.js, Firebase